Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 29 (2012) 4063 – 4067

Procedia Engineering

www.elsevier.com/locate/procedia

2012 International Workshop on Information and Electronics Engineering (IWIEE)

Research on an Application-layer Network Performance Measure Method

Li-jie Cui[a*], Jin-gang Li[a], Wei Liu[b]

[a]*School of Software and Engineering , Harbin University of Science and Technology, Harbin, China*
[b]*Harbin Institute of Technology Software Engineering Co. Ltd, Harbin, China*

Abstract

This paper propose an application-layer network performance measure method for distributed information collection system, and verified the authenticity and stability of the proposed network distance by experiments. Based on the research process, we proposed a definition of network distance based on distributed information collection system of distributed search engine, and to reduce the network of distributed data acquisition system from the total amount of the research for this article, is the main direction. Aimed at the definition of the network distance in information collection system, improved network distance prediction algorithm and Web partition system, experiments and analysis are all carried out. We finish an experimental verification of application-layer network distance that limit the authenticity and stability in statistics less than 10%, through which as a standard of network distance of establishing network coordinate, also can provide a theoretical basis for the further network distance prediction.

Keywords: Network distance; Network performance measure; distributed information retrieval

1. Introduction

Accurate network distance prediction between information collection node and websites distributed in different locations is the basis for collaborative information collection, and also localize the process of collecting sites, thereby reduce the network distance overhead, improve collection efficiency, reduce network load, enhance fault-tolerance capability[1], which is of great significance.

* * Corresponding author.
E-mail address: andyclj1977@163.com

1877-7058 © 2011 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.proeng.2012.01.620

The abstract notion network distance establishes the different metrics for different applications, mainly from the network node connectivity, and communication capability. Current mainstream network distance are the following[2] : 1) round-trip time (RTT): from the beginning of sender sending data to receiving confirmation from the receiving end (the receiver sends an acknowledgment immediately after receiving data) is a total RTT. 2) hop number (Hop): represents the total number of devices transmitted by a specific data (packet). 3) the derived value from RTT and hop number.

The above-mentioned network distance mainly related to network level, without considering the size and number of Web page concurrent downloaded in information collection system, leads to difficult to predict accurately the network distance in the information collection system. Therefore, Section 2 describes related work, and analysis the existing problems. Section 3 and Section 4 propose an application-layer network performance measure method for distributed information collection system, and verified the authenticity and stability of the proposed network distance by experiments.

2. Related work

GNP (Global Network Positioning)[3] was the first to be put forward based on the absolute coordinates of the network distance prediction algorithm, which models the Internet into a geometric space, and corresponds the node in the Internet of any location to the geometric space. In this way, any two nodes of the network are available through the distance between two nodes is modeled to estimate the geometric distance. NPS[4] system further optimized the performance of the network coordinates. PIC system was to optimize the accuracy of the network coordinates, so that it can remove the malicious node limit. LightHouse enhanced the network coordinate system robustness to reduce communication bottlenecks and single points of failure problems.

All these afore-mentioned network distance prediction method applied to distributed information collection system indicates some localization that they are not very good use of the relative location factor in nodes distributed in WAN(wide area network) to improve system efficiency. By considering that the relative location between crawlers in WAN and networking site will affect search engine efficiency. For this reason, this paper propose an application-layer network distance for information collection, to lay foundation for network distance prediction and Web division.

3. Application-layer Network distance for information collection

Information collection system requires crawlers and site nodes network distance to define and measure. Confining websites link from crawlers to the internal cluster is the way to improve the localization ratio of network flow.

In the crawler system, we are more concerned about the time of crawling pages, so network distance in this paper is defined as the time of each crawler crawling each page. Information collection system network distance is defined as follows:

$$Network\_Distance = [(2*RTT) + (L_{page} / BW - 2*RTT)*p] / p \quad (1)$$

RTT is the round-trip time. From the beginning of sender sending data to receiving confirmation from the receiving end (the receiver sends an acknowledgment immediately after receiving data) is a total RTT.

BW is the instantaneous download speeds, that bandwidth.

Went through two TCP/IP handshake, a total time-consuming is 2 * RTT. Download time for P pages is $(L_{page}/BW - 2*RTT)*p$. As a result, the average consumption of each Web page download time is $[(2*RTT) + (L_{page}/BW - 2*RTT)*p]/p$.

4. Application-layer network stability verification

4.1. Experimental Environment

Websites need to be crawled: from the real network, a selection of Chinese provinces and autonomous regions are located in the 203 large sites (including 102 stable node). Allow the system to capture these 203 sites, respectively. Denote the target group for the web site by W = {203}, the number of Nw = 203. Crawler nodes: Redhat Linux server, IP is 202.118.236.137; Network environment: China Education and Research Network Backbone Node; Geographic environment: Harbin Institute of Technology Network and Information Security Experiment Center.

4.2. Experiment content

(1)Experiment design

Experimental method in this paper measured bandwidth value and crawler to a website delay on the condition of all-weather, multi-period, and the relatively long time to verify the final value of network delay and bandwidth values in terms of discrete coefficient stabilized by a large number of experimental data obtained a weighted average value of the delay distribution.

(2)Measurement

Describe RTT value of a crawler with dispersion coefficient. Dispersion coefficient CV (Coefficient Value), also known as the coefficient of variation can be used to describe the dispersion degree of system load, which is the ratio of a standard deviation of a set of data and the corresponding average. The formula is:

$$cv = \frac{s}{\overline{x}} \tag{2}$$

The standard deviation s in formula (5) is the square root of the variance. Standard deviation is dimensional, with the same measure unit of variable with more clearly practical significance. Therefore, the analysis of practical problems use mostly the standard deviation. Standard deviation is calculated as:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} \tag{3}$$

The degree of freedom of the standard deviation in formula (6) is n-1 , because in the calculation of $\sum(x_i - \overline{x})^2$, you must first find the sample average $\overline{x}$, and $\overline{x}$ is attached to a constraint of $\sum(x_i - \overline{x})^2$. Therefore, when calculating the standard deviation, there are only n-1 independent observation values.

Dispersion coefficient is a measure of the relative degree of dispersion data statistics mainly used to compare data for different samples. Dispersion coefficient is large, indicating a large degree of dispersion of data; dispersion coefficient is small, the data shows the degree of dispersion is also small.

Evaluate the network distance prediction result with relative error:

$$\mathrm{Re}\,lativeError = \frac{|PD - AD|}{\min(PD, AD)} \tag{4}$$

PD: Predict Distance, AD: Actual Distance

Predict network distance equal to the actual network distance can be obtained relative error which is 0. The smaller the relative error of prediction showed that the higher accuracy and the better result.

Forecast distance is based on information collected the network node from the definition of distance between the crawler and the website. Actual distance is the test time required to get to the download page.

To get a better visual effect, we used the accuracy to evaluate network distance prediction:

$$\Pr ecision = 1 - \mathrm{Re}lativeError \quad (5)$$

(3)Procedure

1)Web site selection: from the real network, a selection of Chinese provinces and autonomous regions are located in the 203 large sites (including 102 stable node). Allow the system to capture these 203 sites, respectively;2)The target site group is W = {203} sites, Nw = 203;3)Crawler nodes：Redhat Linux server, IP is 202.118.236.137;4) Measure RTT value of the website group with a crawler node per hour;5)Get the download speed of crawling each website home page with crawler node and the corresponding value of the bandwidth BW; 6)Continue the second step to the fifth step, five consecutive days Received a total of 24 * 5 = 120 sets of data;7)Calculate the dispersion coefficient for each node of the target website group; 8)Observe the dispersion coefficient changes, and compare the network distance from the true value and predicted value of the experimental results.

4.3. Experimental Result

Figure 1, Figure 2 shows the distribution of RTT and BW of the five continuous days' measurement per hour.

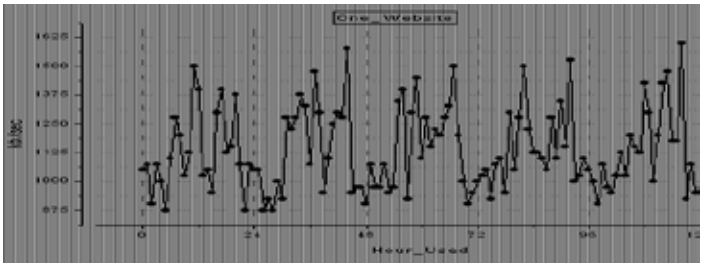

Fig. 1(a) Measured BW distribution of five continuous days (b) Distribution of BW dispersion coefficient

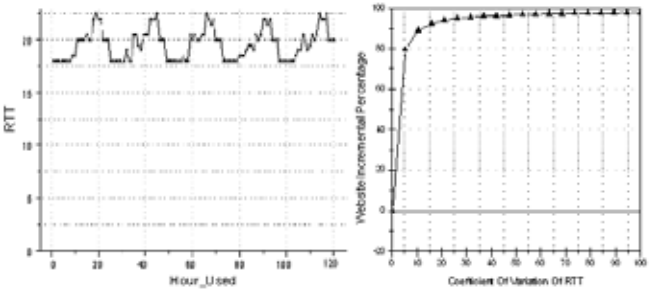

Fig. 2 (a)Measured RTT distribution of five continuous days (b) Coefficient of variate of RTT distribution

Figure 3 shows the comparison figure of network distance value computed by measured BW, RTT and network distance of information systems,and the actual time to download this page(the true network distance value),where the data are measured average . As shown in Figure 3, from the comparison of network distance and measured time used to download web pages (the true network distance),which deduced by information acquisition system and definition of the network distance, the mean relative error of deduced distance and measured distance ranged between 8.26% and 17.36%, and most concentrated in pages which download time more than 500 ms, which may be concerned with large amount of information and the p-value of pipelining, the network dynamics and instability is one of the reasons, either.

According to the results, in the same days of many measurements, changes of BW per hour are relatively dramatic and unstable, but in the same times in different dates BW changes little, while RTT is

a relatively stable volume. Meanwhile, compared with the actual value of network distance, the predicted value has a mean relative error less than 10%,which is acceptable in statistical sense.

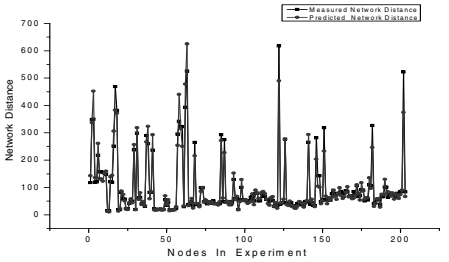

Fig. 3 Contrast of measured network distance and deduced network distance

Therefore, we can take

$$Network\_dis\tan ce = \{2 * \underset{i}{mid}(RTT_i) + ([\overline{L}_{page} / \underset{j}{mid}(BWj)] - 2 * \underset{i}{mid}(RTT_i)) * p\} / p \quad (9)$$

i, j, respectively, correspond to the statistical number of delay and bandwidth. RTT and BW, from the sense of statistical coefficient, tends to stable. That is to say, in the same period every day, RTT and BW of Web sites have small-range fluctuation and high stability, so that make the definition of network distance practical significances which based on RTT and BW. Also, because of the stabilization of these values, a relatively stable network distance value can be acquired, in order which can we establish network coordinates further network prediction system, laying a basis for network distance prediction and Web partition.

5. Conclusion

This paper presents a concept of application-layer network distance for information collection system, and deduces the formal definition of network distance from the theory of web collection and other relevance theories. Finally we finish an experimental verification of application-layer network distance that limit the authenticity and stability in statistics less than 10%, through which as a standard of network distance of establishing network coordinate, also can provide a theoretical basis for the further network distance prediction.

Acknowledgments

This paper was partially supported by the National Natural Science Foundation of China under Grant (No.2010AA012504,NO.2011AA010705); the National Grand Fundamental Research 973 Program of China under Grant (No. 2007CB311101,NO.2011CB302605).

References

[1] Ghemawat S, Gobioff H, Leung S. The Google File System. In: the 19th ACM Symposium on Operating Systems Principles. Google, 2003.

[2] N. R. Sakthivel, V. Sugumaran, Binoy B. Nair Application of Support Vector Machine (SVM) and Proximal Support Vector Machine (PSVM) for fault classification of monoblock centrifugal pump. Dec. 2009: 38-61.

[3] T S Eugene Ng H Z. Towards Global Network Positioning. In: ACM SIGCOMM Internet Measurement Workshop.San Francisco, CA: 2001.